Estimating badger social-group abundance in the Republic of Ireland using cross-validated species distribution modelling

Byrne, Andrew W.[a,b]*, Acevedo, Pelayo[c,d], Green, Stuart[b] and O'Keeffe, James[a,e]

[a] Centre for Veterinary Epidemiology and Risk Analysis (CVERA), School of Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland.

[b] Teagasc Research Centre (Spatial Analysis), Mellows Campus, Athenry, Galway, Ireland.

[c] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBio Laboratório Associado, Universidade do Porto, Portugal

[d] SaBio IREC, Instituto de Investigación en Recursos Cinegéticos (CSIC-UCLM-JCCM), Ciudad Real, Spain.

[e] Department of Agriculture, Food and the Marine (DAFM), Ireland.

* corresponding author: andrew.byrne@ucd.ie; Telephone: +353 (0) 1 7166146.

Word count: 7371

Key words: *Meles meles*, biogeographical model, population size and density estimation, *Mycobacterium bovis*, ecological epidemiology

Abstract

The badger (*Meles meles*) is an important wildlife host for bovine tuberculosis (bTB), and is a reservoir of infection to cattle. Reliable indicators of badger abundance at large spatial scales are important for informing epidemiological investigation. Thus, we aimed to estimate badger social group abundance from a large-scale dataset to provide useful information for the management of bTB in the Republic of Ireland (ROI). Robust estimates of species abundance require planned systematic surveying. This is often unfeasible at large spatial scales, resulting in inadequate (biased) data collection. We employed species distributional modelling (SDM) using 7,724 badger main-sett (burrow) locations across the ROI at a 1ha scale. This dataset was potentially biased as surveying was directed towards areas with cattle bTB-breakdowns. In order to manage sampling bias, we developed a model where the environment was sampled using pseudoabsences geographically constrained to the potential survey area only (constrained model), in addition to a model where all of the ROI was sampled (non-constrained model). Models predictive performance was assessed using internal (splitting the national-scale dataset) and external validation on independent datasets; the latter included 278 main setts from a local-scale unbiased intensive survey (755km^2). Finally, the relationship between predicted probability and observed abundance at local-scale was used to infer number of social-groups at the national level. The geographically constrained model showed moderate discriminatory power, but good calibration in both the internal and external validations. The non-constrained model resulted in higher discrimination but poorer calibration in the internal validation, indicating a limitation for national-scale predictions. Interestingly, there was a strong cubic relationship between predicted probability-classes and observed sett density in the local-area (R^2=0.85 and 0.96; for the non-constrained and the constrained models, respectively). At the national-scale, the preferred model predicted a total of 19,200 (95% Confidence Interval: 12,200-27,900) social groups. Our analyses

42 demonstrated that under a critical perspective large-scale potentially biased datasets can be
43 used to estimate variations in species abundance. The abundance predictions are in keeping
44 with recent independent estimations of the badger population, and will be a valuable index of
45 species abundance for epidemiology (e.g. risk mapping), species management (e.g. informing
46 vaccine strategies) and conservation planning (e.g. assessing population viability).

1. Introduction

Species distribution modelling (SDM) is a rapidly expanding area of research and is quickly becoming an essential tool for studying species distribution ranges and abundances for conservation and wildlife managers (Elith and Leathwick, 2009; Peterson et al., 2011). For instance, SDM is a useful approach to obtain large-scale information of wildlife species that harbour zoonotic infections, which is highly demanded for spatial epidemiology and disease control (e.g. Acevedo et al., 2014a; Ward et al., 2009; White et al., 2008). The advent of widely available large-scale digital datasets (e.g. land-cover, digital elevation models, etc.) as well as the development of robust computational and statistical software to model large datasets are contributing factors to this recent surge in interest and development (McDonald et al., 2013). Datasets on species occurrences are also becoming more accessible and available to researchers through the collection and dissemination of datasets in national or international clearing houses (e.g. Global Biodiversity Information Facility - GBIF), and through the digitisation of museum collections (Peterson et al., 2011). Many of these datasets are collections of – potentially spatially biased – presence points; locations where a species is known to occur at the time of survey. Often however, absence locations are unknown or are poorly estimated (due to the possibility of false negative errors). This has resulted in the development of alternative procedures for modelling species distribution without precise information of absences: the pseudo-absences.

Relevant uncertainties in SDM are associated to absence data, and absences have strong effects on model parameterization and predictions (e.g. Lobo et al., 2010). For these reasons, researchers working with only reliable data for species presence have explored procedures to improve the selection of an appropriate pseudo-absence dataset: randomly (e.g. Wisz and Guisan, 2009), environmentally (e.g. Engler et al., 2004; but see Chefaoui and Lobo, 2008) or spatially stratified selection (Hirzel et al., 2001). Barbet-Massin et al. (2012) concluded that

the suitability of each procedure to select pseudo-absences highly depends of the algorithm used for modelling. For instance, they highlighted that randomly selected pseudo-absences yielded the most reliable models using regression techniques. In addition to the uncertainties in absence data, presence data obtained from opportunistic surveys often exhibit strong spatial bias in survey effort, meaning that some localities are more likely to be surveyed than others (e.g. Reddy and Dávalos, 2003). To address this problem, Phillips et al. (2009) proposed to select pseudo-absences so they reflect the same sample selection bias as the presence data. These authors showed that this procedure produces a more reliable picture of the species range, avoiding the overrepresentation of survey effort, than models developed with randomly selected pseudo-absences.

In this context, we aimed to produce a large spatial scale index of badgers (*Meles meles*) abundance from a potentially biased dataset of main-sett (burrow) occurrence as a relevant tool for the management of bovine tuberculosis (bTB) in the Republic of Ireland (ROI). Badgers are an important wildlife reservoir species for *Mycobacterium bovis*, the causative agent of bTB, in Britain and Ireland, and have been epidemiologically linked with the disease in cattle (Griffin et al., 2005). Wildlife abundance estimates in these situations are highly demanded, especially where high profile wildlife conflicts are apparent (Acevedo et al. 2014a). Some potential bias was expected in our dataset since sett occurrences were obtained under a survey exclusively motivated by an epidemiological investigation into the potential causes of cattle herd bTB breakdown, that is, a design far from ideal when determining the badger density distribution in ROI. Main sett numbers can be used as a proxy for social group abundance, as has been used frequently elsewhere (e.g. Acevedo et al., 2014b; Judge et al., 2014), and was shown to be a good indicator of badger abundance (e.g. Lara-Romero et al., 2012), therefore modelling setts occurrence an index of badgers relative abundance can be also estimated. However, some caution has to be employed when extrapolating to badger

abundance due to the variation in social group sizes and the rare occurrence of two main setts within one territory (Byrne et al., 2012a).

Badger-habitat relationships are relatively well known in Ireland and Britain (e.g. Hammond et al., 2001; Newton-Cross et al., 2007) and they were used to predict badger or sett abundance in these regions previously (Etherington et al., 2009; Reid et al., 2012; Sleeman et al., 2009). However, little is known in ROI about the spatial variation the badger density, especially in areas that do not have bTB problems in cattle, and there are currently no large-scales indices of badger abundance in this country. This study is the first to predict the spatial variation in badger social group density at the national scale in ROI. The current study benefits from having both an extensive large-scale dataset (30,610 setts; 7,724 main setts) and a smaller-scale intensively surveyed dataset (1,009 setts; 278 main setts) from which internal and external validation processes of the model predictions can be implemented. The results of this study are particularly important for future epidemiological modelling and for the design of disease management strategies.

2. Methods

2.1 Datasets

Two datasets were utilised in the present study – each collected for different purposes, at different scales and survey intensity. Both datasets include the presence of main and non-main setts. Main setts are large burrow systems; larger than non-main setts and more frequently used by the badgers of a social group (Byrne et al., 2012a). Typically there is one main sett per social group; therefore the number of main setts can be used as a proxy for badger social group abundance (Byrne et al., 2012b; Etherington et al., 2009; Judge et al., 2014). Main setts are conspicuous and exist for long periods of time (>100 years in some recorded cases; Byrne et al., 2012a), and so they are likely not to suffer greatly from detectability issues. The first dataset ("national-scale dataset") was generated due to a national-scale badger management program where badger sett surveys are required as part of an epidemiological investigation into the potential causes of cattle herd bTB breakdowns (Byrne et al., 2013a, b). These location data were generated during 2004-2012. The second dataset ("local-scale dataset") was generated during a vaccination trial in north-west Co. Kilkenny (see Byrne et al., 2012b). The area was surveyed intensively during a mark-recapture study from 2008-2012. The survey area was delineated as part of the study design with borders composing primarily of rivers and roads.

2.2 National-scale Dataset

During these surveys, field staff record the location of setts found on index herd (the bTB breakdown farm) farm land and in areas up to 2km from the index herd land parcel boundary. The greatest survey intensity is focused on the index farm land and the contiguous herds immediately surrounding the index herd. However, focal surveys of the surrounding

134 landscape are undertaken; this includes the use of detailed maps and orthophotography to
135 locate and survey areas with greater likelihood of badger sett presence. Local farmer and
136 huntsman knowledge supplement the surveying effort. Because of this approach, we are
137 certain of the location of setts, but we are uncertain as to the true extent of the survey and
138 where true absences can be located. As the surveying of badgers was directed towards areas
139 with herd bTB breakdowns, we were aware that there is a possible sampling bias within the
140 dataset. We implemented a geographic constraint by choosing pseudo-absences from areas
141 with the same underlying biased sampling distribution as the presences, in order to address
142 potential sampling bias (Phillips et al., 2009). Fortunately, we know that the maximum
143 distance away from a herd breakdown that was surveyed was 2km, therefore we could use
144 this rule to construct a *potential maximum extent surface* (PMES) from which environmental
145 availability could be assessed (Figure 1). Constraining the pseudo-absences samples from this
146 distribution allows for improved modelling performance while increasing the likelihood of
147 the pseudo-absences representing true-absences (Phillips et al., 2009; Zaniewski et al., 2002).
148 As most of our surveying took place at mid to low latitudes, we also constrained our PMES to
149 areas below 300m ASL (i.e. higher altitudes were undersampled) in order to avoid the
150 inclusion of pseudo-absences beyond the environmental domain represented in the survey.
151 We used 10,000 pseudo-absences to assess available environmental conditions in the
152 sampling area (see Baret-Massin et al., 2012; Phillips and Dudik, 2008; Wisz and Guisan,
153 2009). Pseudo-absence points were restricted from raster cells containing any presences
154 (main and non-main setts; total known setts: 30,610) and a minimum distance between
155 pseudo-absences was imposed (500m coinciding with typical Irish main sett spacing; Byrne
156 et al., 2013b). The spatial scale of the raster dataset was 1ha (our spatial unit for modelling),
157 as most national-scale environmental layers could be scaled to this size and this size was used
158 in a recent badger study for England and Wales (Etherington et al., 2009).

159 To evaluate the PMES constrained approach, we also constructed models for comparison
160 using pseudo-absence locations (10,000) drawn randomly from the entire country (ROI), with
161 the exclusion of the 'local-scale dataset'.

162 **2.3 Local-scale dataset**

163 Extensive surveys, by trained experienced field staff, were undertaken in a 755km^2 area in
164 north-west Co. Kilkenny as part of a bTB vaccination trial for badgers (see Byrne et al.,
165 2012b; Figure 1). Due to the intensive nature of the surveys and the defined boundary
166 delineating the survey extent, these data could be considered a presence-absence dataset
167 (though there may be a low probability that some main setts were undetected or misclassified
168 as non-main setts).

169 **2.4 Modelling approach**

170 We modelled probability of occurrence (see Acevedo and Real, 2012) for badger main sett
171 construction using a binary logistic regression. Logistic models were built by first assessing
172 the relationships between outcome and independent variables using univariate models. All
173 independent variables with significant associations with the outcome variable at α=0.1 were
174 further investigated within a multivarible logistic regression. The predictors used, and their
175 sources, are listed in Table 1. Layers were converted to a raster grid and resampled to a 1ha
176 scale. Coarse grained datasets (e.g. CORINE) formed an index by enumerating the number of
177 25m grids (0.0625ha) of the habitat type within 300m of the 1ha grid square (following Reid
178 et al., 2012). Finer resolution datasets applied the same index, but restricted the search
179 window to 100m (i.e. forest cover). Hedgerow density was measured as an index based on a
180 national map of all hedgerows (vegetated field boundaries) ≥2m in width based upon
181 automated image processing of orthophotography (S. Green, Teagasc).
182 For strongly correlated predictor variables, only the variable most strongly correlated with the

183 outcome variable was retained (Newton-Cross et al., 2007). Screening for linearity was
184 assessed visually using LOWESS smoothed curves. Where non-linear relationships were
185 found dependent variables were suitably transformed (e.g. by the inclusion of quadratic
186 terms; Dohoo et al., 2009; Chapter 15, p365-380). If additional variables (e.g. quadratic term)
187 were significant, they were retained in the model (Dohoo et al., 2009; Chapter 15, p365-380).
188 These quadratic term variables were also centred to decrease the Variance Inflation Factor
189 (VIF) within the model (Dohoo et al., 2009; Chapter 14, p. 338-340). The VIF cut point was
190 ≤10 before variable was centred. We assessed the robustness of our final models using
191 bootstrap analysis (using the SWBOOT function in Stata 10) with 100 bootstrapped repeats
192 of backward stepwise logistic regression on candidate variables (Austin and Tu, 2004).
193 Variables that were significant in >70% of bootstrap samples were included in the final model
194 (Austin and Tu, 2004), with the exception of the geographical coordinates (*x* and *y*) which
195 were retained in all candidate models.
196 Newton-Cross et al. (2007) highlighted how validation of badger-habitat models have been
197 extremely limited, therefore we implemented both internal (splitting the national dataset) and
198 external validation (independent local dataset). National-scale sett data was split into training
199 (70%) and validation (30%) datasets in order to perform an internal cross-validation process.
200 Models were built from the training dataset and then predictions were made on the validation
201 data to assess predictive performance (Fielding and Bell, 1997). As this process is affected by
202 the subsample taken for the training and testing datasets, we repeated the process 10 times
203 (Etherington et al., 2009) and report the mean values of the statistical parameters describing
204 model predictive performance. In addition, an external validation was carried out on the
205 local-scale dataset. We followed Steyerberg et al. (2010) by assessing the predictions of the
206 overall final national models (“development models” based on 7,724 sett locations) on the
207 independent local-scale dataset. For both internal and external validation we evaluated the

discriminatory performance of the models using the area under the ROC curve (AUC; see Lobo et al., 2008). We also estimated Cohen's Kappa, sensitivity (SE), specificity (SP) and the True Skill Statistic as complementary measures of discrimination. The values of the latter statistics are dependent on the threshold chosen, we chose "optimal" thresholds that maximised the highest combination of SE and SP (Reichenheim, 2002). Hosmer-Lemeshow tests were used to evaluate the models in terms of calibration (reliability), as a complementary and informative characteristic of the model's predictive performance (Jiménez-Valverde et al., 2013). We extended our evaluations of calibration on the local-scale dataset (external validation) by assessing the maximum and mean difference in predicted and observed probabilities (E_{max} and E_{avg}, respectively) and the calibration slope (Hosmer and Lemeshow, 2000; Steyerberg et al., 2010). The calibration slope (β) was estimated from a linear regression (predicted probabilities versus observed frequencies), with $\beta \approx 1$ and $\alpha \approx 0$ indicating near perfect calibration (α is the intercept of the calibration slope). Values $\beta < 1$ indicate overfitting; $\alpha \neq 0$ indicates whether predictions are systematically above or below expectation. Inspection of the calibration was initially undertaken by graphing the predicted and observed probabilities using decile bins (a Hosmer-Lemeshow plot) and then using a LOWESS smoothing algorithm (bandwidth: 0.2) against predicted probabilities (Steyerberg et al., 2010); E_{avg} and E_{max} were calculated using the predicted values from the LOWESS smoothing algorithm.

2.5 Assessing the relationship between probability of sett occurrence and social group abundance

We evaluated ability of the national models (constrained and non-constrained) to predict main sett densities by regressing the observed (main sett) density by the predicted probability (Boyce et al., 2002). We classified probability values using 10 quantiles (deciles), and

accordingly calculated the area adjusted density for each class within the Kilkenny area. We fitted simple linear regression models, and investigated 2^{nd} and 3^{rd} order polynomials transformations of the independent variable (i.e. linear, quadratic and cubic transformations) using R^2 as the assessment of model fit (Etherington et al., 2009). We used the best fitting regression model to predict the number of main setts within the Kilkenny area, and compared the result with the observed number of main setts. We used the relationship between probability quantiles and density to map the predicted densities across ROI.

All data manipulation and modelling was performed in Stata 11 (Stata Corp.), and GIS operations were performed in ArcGIS (ESRI).

3. Results

3.1 Factors affecting badger sett occurrence

A number of factors were found associated with badger sett presence across the two modelling approaches, constrained and non-constrained models (see Table S1 and S2 in the Supplementary Material, respectively). Across the two models, setts were most strongly positively affected by local hedgerow density. The relationship between slope, and elevation, and the probability of sett presence was quadratic in nature. This indicates greater likelihood of setts occurring on gentle slopes (<15º) and in moderate altitudes (30m-170m; Figure S1). The constrained model also suggested that there was a significant negative relationship between the sine (eastness) of the aspect of the slope and the probability of sett presence. Badger setts tended to be significantly positively associated with greater forest cover and pasture habitats within close proximity (300m of sett grid square). Setts were negatively associated with blanket bog, water edge and altered man-made surfaces (e.g. open mines and landfills). Badger setts were negatively associated with shallow, poorly drained soil types (constrained model) and positively associated with deep, well-drained soil types (non-constrained model).

3.2 Internal cross-validation

3.2.1 Constrained model

The total area of the PMES was 49,002 km^2 (66% of the total land area of ROI; Figure 1). The constrained national model exhibited moderate discriminatory power, with a mean AUC = 0.71 (range: 0.69-0.73; Table S3) in the internal validation. Other metrics of discrimination are presented in Table S3. The model performed well in terms of calibration in the internal validation datasets (Hosmer-Lemeshow test: mean P = 0.380,

range: 0.160 -0.645; Table S3).

3.2.2 Non-constrained model

The final non-constrained model performed better in comparison with the constrained model in terms of overall discriminatory power, with a mean testing AUC = 0.77 (range: 0.75-0.78), see also Table S4. The model performed poorly in terms of calibration, with a Hosmer-Lemeshow test mean $P = 0.042$ (range: $P<0.001$- 0.200) (see Table S4). Observing the calibration slope, suggested that the model poorly predicted the proportions of presences at high predicted values (i.e. the final decile).

3.3 External validation – local-scale dataset

3.3.1 Constrained model

The final constrained model performed well when tested on the local-scale dataset. There was no drop in AUC between the final constrained model and local-scale dataset, in fact the model performed marginally better at local-scales (internal validation AUC: 0.71 vs. external validation AUC: 0.72; Table 2). When the final model was used to predict to the local-scale dataset, there was some evidence that calibration fit had (non-significantly) declined (Homser-Lemeshow test: $P = 0.087$). The mean difference between observed frequencies and predicted probabilities, E_{avg}, for the final constrained model was -0.0008, whereas it was -0.04 when predicting to the local-scale dataset. However, the β value from the calibration slope for both the final constrained model and the local-scale data were close to 1 (1.001, 1.009 respectively; R^2 =0.97; Table 2), indicating that the model did not overfit the data. The α regression values (both models: -0.05) however, suggested that there may have been some minor systematic negative bias in the model predictions.

3.3.2 Non-constrained model

The final non-constrained model exhibited a decline in discriminatory power when predicted to the independent dataset (internal validation AUC= 0.77 vs. external validation AUC=0.73; Table 2). Despite the final non-constrained model exhibiting problems with regards calibration in the internal validation, there was no significant lack-of-fit when predicted to the local-scale dataset (P=0.574). The calibration slope was <1, indicating that there may have been some minor overfitting; α values were <0, indicating some minor systematic negative bias (Table 2).

3.4 Relationship between probability of sett occurrence and sett density

The final constrained model predictions (probability classes) were best associated with sett density with a cubic regression (R^2 = 0.96; Figure 2b). The model performed very well when predicting main sett numbers, with the mean predicted number of setts (278 main setts; 95% CI: 199 - 363) being the same as observed number of setts within the area (278 main setts). The model predicted 19,159 main setts (95% CI: 12,221-27,898; Table 4; Figure 3), and therefore social groups, at national scale.

The non-constrained model could predict main sett density within the Kilkenny test area well, with the simplest linear regression model that best fitted the data being a cubic regression without a linear trend (R^2 = 0.85; Figure 2a). The model predicted the number of main setts within the Kilkenny test area accurately (278 main setts; 95% CI: 191-372). This non-constrained national model predicted a total number of social groups for all ROI of 17,324 (95% CI: 10,220-25,902; Table 3; Figure 3).

4. Discussion

4.1 Factors affecting sett distribution

Different factors affected badger sett presence depending on the dataset used to construct the models. However, there was a general trend for badger setts in both models to occur in moderately steep areas, at relatively low elevations, in deep, well-drained soils types in areas that had sources of cover (hedgerows and/or forests) and forage (e.g. pasture). These findings concur generally with previous badger-habitat models, generated with data from various scales in Ireland, Britain and in other areas of Western Europe (e.g. Etherington et al., 2009; Hammond et al., 2001; Newton-Cross et al., 2007; Reid et al., 2012; Schley et al., 2004).

4.2 Model performance and general evaluation

Models that attempt to encapsulate the species-habitat relationship of common, widely-distributed, generalist species will generally have poor discriminatory power in comparison with restricted specialist species (McPherson and Jetz, 2007; Evangelista et al., 2008 but see Lobo *et al.* 2008). It has been shown that discriminatory power of a model may be improved by extending the geographic scale of the analysis (Acevedo et al., 2012; Jiménez-Valverde et al., 2013) – which is the complete opposite to what was employed during the present study to deal with the potentially biased nature of our sampling (during the collection of the national dataset) by geographically constraining where pseudo-absence points could be located (Philips et al., 2009; Zaniewski et al., 2002). This was demonstrated by our non-constrained model having greater AUC (0.77) than the geographically constrained model (AUC= 0.71). In terms of presence/pseudo-absence modelling, it is well recognized that an upper AUC limit <1 can only be achieved (Phillips and Dudik, 2008) with perfectly calibrated models theoretically only

reaching an AUC of 0.83 (Jiménez-Valverde et al., 2013). In other fields, it is well recognised that both measures of discrimination and calibration needs to be assessed and reported (e.g. epidemiology, Steyerberg et al., 2010), with calibration being of particular importance for predictive models (see also Jiménez-Valverde et al., 2013). In this context, our models performed well in calibration (during internal and external validation procedures), and is therefore a useful framework to base density estimates. Our use of a number of calibration assessments (e.g. calibration slope (β), error around the slope (E_{avg}), and the calibration intercept (α)) should be considered for future biogeographical studies, especially for generalist species where calibration is important.

Others have reported that data quality is probably the most important factor influencing general model performance, an aspect to which much more effort and resources should be devoted (Feeley and Silman, 2011). Indeed, our models can only be as good as the national-scale independent predictors on which we can base our inferences. While the datasets used in the present study are the best available, greater development and ground-truthing of these layers will undoubtedly improve model performance in the future. Furthermore, there is room to improve the external validation of this model when other datasets become available (due to continuing research on badger densities in Ireland, for example), allowing for the model to become a dynamic tool for conservation and wildlife managers (Byrne, 2013). Such dynamical analytical approaches may be useful for other spatial modelling of wildlife in other territories and/or species (Acevedo et al., 2014b).

4.3 Density estimates at local and national-scales

The probability values predicted from the models were strongly related to the observed main sett densities in the local-scale dataset. These models could predict observed main sett densities well, albeit with relatively wide confidence intervals. The models predicted a national social-group population (based on one main sett per social group; Byrne et al., 2012b; Etherington et al., 2009; Reid et al., 2012) of 17,300 (95% CI: 10,200-25,900) and 19,200 (95% CI: 12,200-27,900) using the non-constrained and constrained models respectively. The latter estimate (the constrained model) is considered the preferred model, as we attempted to address the potential bias in our dataset during this analysis by geographically constraining the selection of pseudo-absences (Phillips et al. 2009). The poor calibration of the non-constrained national model was partially due to a reduced ability of the model to predict the proportions in the last decile of the calibration plot, suggesting the model could underestimate the greatest probability categories. This calibration issue is likely due to the underlying sampling bias introduced during the data collection. However, the non-constrained model performed well at the local-scale, indicating that the sampling bias is likely spatially structured with stronger effects expected in some areas, but not in the local-scale area.

The last national badger population suggested that there were approximately 84,000 (95% CI 72,000 -95,000) badgers present in Ireland (Sleeman et al., 2009). Using the crude measure of mean group size (4.1) reported by Byrne et al. (2012b), this population estimate would suggest that the population was composed of 20,500 (95% CI 17,600-23,200) groups. This suggests that the estimations made during the present study are plausible.

Our density estimates for each probability class were remarkably similar to previous work done in Northern Ireland, England and Wales (Etherington et al., 2009; Feore, 1994; Reid et al., 2012). Both Feore (1994) and Reid et al. (2012) found that social group density in Northern Ireland varied significantly amongst landscape classes with poor habitats (e.g. mountains) having a mean density of 0.08-0.33km^{-2}, while optimum habitats (e.g. drumlin farmland) having a mean density of 0.81-0.85km^{-2}. Overall, the present study suggests that mean group density in the ROI is substantially lower than in Northern Ireland (Figure 4a). This is mainly due to the fact that a large area of the ROI is of poor favourability for badgers (Figure 4b; Figure 3). Bog lands in central and western Ireland (i.e. raised and blanket bog habitats), wet lands of the north-west and mountainous areas in the south-west make up significant areas of low badger social group density. The estimated densities across classes are similar to those estimated for England and Wales (Figure 4b). Across all four jurisdictions, the mean estimated densities for highly suitable landscapes are remarkably similar (range: 0.81-1.22) despite methodological differences between studies (Figure 4b).

4.4 Practical applications – Indicator for epidemiology, management and conservation

Indices of badger abundance have been associated with increasing risk of bTB in cattle herd breakdowns in both Britain and Ireland (e.g. Bessell et al., 2012; Griffin et al., 2005). A ‘risk-map’ of badger abundance has been used to assess the relationship between indices of badger abundance nationally and likelihood of herd breakdowns in Britain (Ward et al. 2009; White et al. 2008). The resolution of the maps produced in this study will allow for future fine resolution spatial analyses of the badger-cattle epidemiological relationship within the ROI. Furthermore, the model outcomes presented here will be essential for assessing the potential risk to farms outside of the core bTB

areas from which the raw data was derived. Badger vaccination programmes (intra-muscular, Bacille Calmette-Guérin) are currently being developed in six large-scale project areas in the ROI (2,796km^2 total area) as a potential means of controlling bTB in badgers and, ultimately, in cattle populations (O'Keeffe, J., Byrne, A. and Martin, S.W., unpublished). Designing, and implementing, such programs require indicative abundance maps such as those produced during this analysis to inform planning and management to maximise vaccine coverage.

Badgers populations in ROI are currently managed (culled) on approximately 30% of the agricultural land (Byrne et al., 2013a; Sheridan, 2011). This regime has significantly decreased the relative abundance of badgers locally (Byrne et al., 2013b). The present models will facilitate managers and conservationists to design future programs that ensure the potential disease benefits of population control (Griffin et al., 2005), while ensuring the long-term viability of badger's populations, as required by international legislation (Byrne, 2013).

Conclusion

We have shown that large-scale opportunistic datasets with reliable presence data can be used to capture the underlining structure of the species-environment relationship from spatially explicit models. The models, though based on data collected for other purposes, performed well as a tool to estimate probability of sett occurrence and abundance. Internal and external validations suggested that the models were well calibrated and had the ability to predict on independent datasets without large drops in performance. Predictions of sett density based on these models were in-line with previous work suggesting the results are plausible. As badgers are of particular interest due to their role in the epidemiology of *M. bovis*, the causative agent of bTB, the distribution and density

425 estimates from these models will be of particular utility for future epidemiologic
426 research, and will form the basis of wildlife disease reservoir ‘risk-mapping’.
427 Furthermore, as badger populations are currently under a culling management regime in
428 part of the badger distribution in ROI, these estimates will be a cornerstone for future
429 conservation assessments investigating the impact of culling on the national badger
430 population.

431 **Acknowledgments**

432 The authors wish to acknowledge and thank G. McGrath (CVERA) for extracting the data
433 from the national GIS raster files (7 million records) and P. White (DAFM/CVERA) for
434 advice with database extraction procedures and discussions on the modelling approaches.
435 AWB was funded by a Teagasc Walsh Fellowship (www.teagasc.ie) and a Post-doctoral
436 Research Fellowship (PDRF-L1) within the Centre for Veterinary Epidemiology and Risk
437 Analysis (www.ucd.ie/cvera). P. Acevedo enjoyed a post-doctoral grant
438 (SFRH/BPD/90320/2012) from 'Fundação para a Ciência e a Tecnologia' (FCT) funded
439 by 'Programa Operacional Potencial Humano' (POPH) – 'Quadro de Referência
440 Estratégico Nacional' (QREN) from the European Social Fund and by the Portuguese
441 'Ministério da Educação e Ciência'. He is currently supported by the 'Spanish Ministerio
442 de Economía y Competitividad' (MINECO) and 'Universidad de Castilla-La Mancha'
443 (UCLM) through a ‘Ramón y Cajal’ contract (RYC-2012-11970) and partly by EMIDA
444 ERA-NET grant Aphaea (219235 FP7 ERA-NET EMIDA; www.aphaea.eu). All data
445 collection and database maintenance was funded by the Irish Department of Agriculture,
446 Ireland (www.agriculture.gov.ie).

References

Acevedo, P., Quirós-Fernández, F., Casal, J., Vicente, J., 2014a. Spatial distribution of wild boar population abundance: Basic information for spatial epidemiology and wildlife management. Ecol. Indicators, 36, 594-600.

Acevedo, P., González-Quirós, P., Prieto, J. M., Etherington, T. R., Gortázar, C., Balseiro, A. 2014b. Generalizing and transferring spatial models: A case study to predict Eurasian badger abundance in Atlantic Spain. Ecol. Model., 275, 1-8.

Acevedo, P., Real, R., 2012. Favourability: concept, distinctive characteristics and potential usefulness. Naturwissenschaften, 99, 515–522.

Acevedo, P., Jiménez-Valverde, A., Lobo, J.M., Real, R. 2012. Delimiting the geographical background in species distribution modelling. J Biogeog. 39, 1383–1390.

Austin, P.C., Tu, J.V., 2004. Bootstrap methods for developing predictive models. Am. Stat. 58, 131–137.

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3: 327-338

Bessell, P. R., Orton, R., White, P. C., Hutchings, M. R., Kao, R. R., 2012. Risk factors for bovine Tuberculosis at the national level in Great Britain. BMC Vet. Res., 8, 51.

Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K.A., 2002. Evaluating resource selection functions. Ecol. Model. 157, 281–300

Byrne, A.W., 2013. Studies relating to the population dynamics of the European badger (*Meles meles*) in Ireland. PhD Thesis, University College Cork, Ireland.

468 Byrne, A.W., O'Keeffe, J., Green, S., Sleeman, D.P., Corner, L.A.L., *et al.*, 2012b.
469 Population estimation and trappability of the European Badger (*Meles meles*):
470 Implications for tuberculosis management. PLoS ONE, 7, e50807
471 doi:101371/journalpone0050807

472 Byrne, A.W., O'Keeffe, J., Sleeman, D.P., Davenport, J., Martin, S.W., 2013a. Factors
473 affecting European badger (*Meles meles*) capture numbers in one county in Ireland. Prev.
474 Vet. Med. 109, 128-135.

475 Byrne, A.W., O'Keeffe, J., Sleeman, D.P., Davenport, J., Martin, S.W. 2013b. Impact of
476 culling on relative abundance of the European badger (*Meles meles*) in Ireland. Eur. J.
477 Wildl. Res. 59, 25-37.

478 Byrne, A.W., O'Keeffe, J., Sleeman, D.P., Davenport, J., 2012a. The ecology of the
479 European badger (*Meles meles*) in Ireland – a review. Biol. Environ. 112, 105–132.

480 Chefaoui, R., Lobo, J.M., 2008. Assessing the effects of pseudo-absences on predictive
481 distribution model performance. Ecol. Model. 210, 478-486

482 Dohoo, I.R., Martin, S.W., Stryhn, H. 2009. Veterinary epidemiologic research, 2nd edn.
483 VER Inc, Canada.

484 Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and
485 prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677-697.

486 Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the
487 distribution of rare and endangered species from occurrence and pseudo-absence data. J.
488 Appl. Ecol. 41: 263–274

489 Etherington, T.R., Ward, A.I., Smith, G.C., Pietravalle, S., Wilson, G.J., 2009. Using the
490 Mahalanobis distance statistic with unplanned presence-only survey data for
491 biogeographical models of species distribution and abundance: a case study of badger
492 setts. J. Biogeog. 36, 845–853.

493 Evangelista, P.H., Kumar, S., Stohlgren, T.J., Jarnevich, C.S., Crall, A.W., Norman III,
494 J.B., Barnett, D.T. 2008. Modelling invasion for a habitat generalist and a specialist plant
495 species. Diversity and Distributions, 14: 808–817.

496 Feeley, K.J., Silman, M.R., 2011. Keep collecting: accurate species distribution
497 modelling requires more collections than previously thought. Divers. Distrib. 17, 1132–
498 1140.

499 Feore, S.M., 1994. The distribution and abundance of the badger *Meles meles* L in
500 Northern Ireland. PhD thesis, Queen's University of Belfast, United Kingdom.

501 Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction
502 errors in conservation presence/absence models. Environ. Cons. 24, 38–49.

503 Griffin, J.M., Williams, D.H., Kelly, G.E., Clegg, T.A., O'Boyle, I., Collins, J.D., More,
504 S.J., 2005. The impact of badger removal on the control of tuberculosis in cattle herds in
505 Ireland. Prev. Vet. Med. 67, 237-266.

506 Hammond, R.F., McGrath, G., Martin, S.W., 2001. Irish soil and land-use classifications
507 as predictors of numbers of badgers and badger setts. Prev. Vet. Med. 51, 137-148.

508 Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a
509 virtual species. Ecol. Model. 145, 111–121

510 Hosmer, D.W., Lemeshow, S., 2000. Applied logistic regression, 2nd edn Wiley, New
511 York.

512 Jiménez-Valverde, A., Acevedo, P., Barbosa, A.M., Lobo, J.M., Real, R., 2013.
513 Discrimination capacity in species distribution models depends on the representativeness
514 of the environmental domain. Global Ecol. Biogeog. 22, 508–516.

515 Judge, J., Wilson, G.J., Macarthur, R., Delahay, R.J., McDonald. R.A. 2014. Density and
516 abundance of badger social groups in England and Wales in 2011–2013. Sci. Rep. 4,
517 3809. doi:10.1038/srep03809.

518 Lobo, JM, Jiménez-Valverde, A Real, R, 2008. AUC: a misleading measure of the
519 performance of predictive distribution models Global Ecol. Biogeog. 17, 145–151.

520 Lobo, JM, Jimenez-Valverde, A, Hortal, J., 2010. The uncertain nature of absences and
521 their importance in species distribution modelling. Ecography, 33, 103-114.

522 McDonald, L., Manly, B., Huettmann, F., Thogmartin, W., 2013. Location-only and use-
523 availability data: analysis methods converge. J. Anim. Ecol. 82, 1120–1124.

524 McPherson, J.M., Jetz, W., 2007. Effects of species' ecology on the accuracy of
525 distribution models. Ecography, 30, 135–151

526 Murphy, D., Gormley, E., Collins, D. M., McGrath, G., Sovsic, E., Costello, E., Corner,
527 L.A.L., 2011. Tuberculosis in cattle herds are sentinels for *Mycobacterium bovis*
528 infection in European badgers (*Meles meles*): The Irish Greenfield Study. Vet. Microbiol.
529 151, 120-125.

530 Newton-Cross, G., White, P.C.L., Harris, S., 2007. Modelling the distribution of badgers
531 *Meles meles*: comparing predictions from field-based and remotely derived habitat data.
532 Mamm. Rev. 37, 54–70.

533 Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E.,
534 Nakamura, M., Araújo, M,B., 2011. Ecological niches and geographic distributions.
535 Princeton University Press, USA.

536 Pfeiffer, D.U., 2009. Analysis of spatial data. Chapter 26. In I.R. Dohoo, W. Martin and
537 H. Stryhn (eds): Veterinary epidemiological research. 2nd ed. AVC Inc.

538 Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of
539 species geographic distributions. Ecol. Model. 190, 231–259.

540 Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new
541 extensions and a comprehensive evaluation. Ecography, 31, 161–175.

542 Reddy, S., L.M. Dávalos, 2003. Geographical sampling bias and its implications for
543 conservation priorities in Africa. J. Biogeog. 30,1719–1727.

544 Reichenheim, M.E., 2002. Two-graph receiver operating characteristic. Stata J. 2, 351–
545 357.

546 Reid, N., Etherington, T.R., Wilson, G.J., Montgomery, W.I., McDonald, R.A., 2012.
547 Monitoring and population estimation of the European badger *Meles meles* in Northern
548 Ireland. Wildl. Biol. 18, 46-57.

549 Schley, L., Schaul, M., Roper, T. J., 2004. Distribution and population density of badgers
550 *Meles meles* in Luxembourg. Mamm. Rev. 34, 233–240.

551 Sleeman, D.P., Davenport, J., More, S.J., Clegg, T.A., Collins, J.D., Martin, S.W.,
552 O'Boyle, I., 2009. How many Eurasian badgers *Meles meles* L are there in the Republic
553 of Ireland? Eur. J. Wildl. Res. 55, 333-344.

554 Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N.,
555 Kattan, M.W., 2010. Assessing the performance of prediction models: a framework for
556 some traditional and novel measures. Epidem. 21, 128-138.

557 Ward, A.I., Smith, G.C., Etherington, T.R., Delahay, R.J. 2009. Estimating the risk of
558 cattle exposure to tuberculosis posed by wild deer relative to badgers in England and
559 Wales. J. Wildl. Dis. 45, 1104-1120

560 White, P.C., Böhm, M., Marion, G., & Hutchings, M. R., 2008. Control of bovine
561 tuberculosis in British livestock: there is no 'silver bullet'. Trends Microbiol. 16, 420-
562 427.

563 Wisz, M.S., Guisan, A., 2009. Do pseudo-absence selection strategies influence species
564 distribution models and their predictions? An information-theoretic approach based on
565 simulated data. BMC Ecol., 9, 8.

566 Zaniewski, A.E., Lehmann, A., Overton, J.M., 2002. Predicting species spatial
567 distributions using presence-only data: a case study of native New Zealand ferns. Ecol.
568 model. 157, 261-280.
569

570 **Supporting Information**

571 Additional Supporting Information may be found in the online version of this article:

572 **Appendix S1** Model summaries and internal validation.

573 **Appendix S2** Relationship between altitude/slope and sett-presence.

574

575 **Table 1**

576 Descriptions of independent variables used to construct habitat models for badgers in the
577 Republic of Ireland. These predictors were chosen as they were found to be significantly
578 associated with badger distribution in previous studies (see text).

Independent variables	Description	Derived
Aspect (sine and cosine)	Continuous; Index	Digital Elevation Model (20m resolution)
Slope (degrees)	Continuous; Degree	Digital Elevation Model (20m resolution)
Elevation (centred with quadratic term)	Continuous; Metres	Digital Elevation Model (20m resolution)
TOPEX	Continuous; Index of topographical exposure	TOPEX model (Teagasc; S. Green)
Geographic coordinates	Continuous; Metres	Irish Grid (Transverse mercator)
Pasture	Continuous; Number of 25m raster cells within 300m	CORINE (level 3)
Arable	Continuous; Number of 25m raster cells within 300m	CORINE (level 3)
Hedgerow	Continuous; Index (0-8853)	Teagasc hedgerow map (Teagasc; S. Green)
Forest cover	Continuous; Number of 25m raster cells within 100m	Forest Cover Map (DAFF 2007)
Soil type	Categorical/dichotomous	Teagasc EPA soils and subsoils (Teagasc; S. Green)
Parent material	Categorical/dichotomous	Teagasc EPA soils and subsoils (Teagasc; S. Green)
Distance to river/lake	Continuous; Meters	National digital map of lakes and rivers (DAFM/CVERA)

579
580

Table 2

Performance of the national constrained and non-constrained badger sett distribution models in the Republic of Ireland using the internal validation (independent subset of data at national-scale) and the external validation (local-scale) dataset.

	Constrained				**Non-constrained**			
	Internal validation		**External validation**		**Internal validation**		**External validation**	
AUC	0.71		0.72		0.77		0.73	
TSS	0.31		0.35		0.39		0.39	
Sensitivity	64.18		66.55		70.53		71.58	
Specificity	66.48		68.32		68.84		67.81	
Kappa	0.33		0.33		0.33		0.32	
E_{avg}	-0.001	(R^2: 0.97)	-0.04	(R^2: 0.97)	-0.001	(R^2: 0.98)	-0.04	(R^2: 0.98)
E_{max}	-0.15		0.15		-0.17		-0.26	
β	1.001		1.009		0.988		0.994	
α	-0.05		-0.05		-0.04		-0.05	
HL-gof	0.256		0.087		<0.001		0.574	

AUC: Area Under the ROC Curve; TSS: True Skill Statistic; E_{avg}: Mean difference in observed and predicted probabilities; E_{max}: Maximum difference in observed and predicted probabilities; β: slope of the calibration regression; α: intercept of the calibration regression; HL-gof: Hosmer-Lemeshow goodness of fit test.

590 **Table 3**

591 Badger social group predictions of the non-constrained national model (pseudo-absences

592 taken from across the country) for the Republic of Ireland.

Probability class (deciles)	Estimated density	Estimate lower density	Estimate upper density	Km^2	Total estimate	Lower 95% CI	Upper 95% CI
1	0.081	0	0.205	26,760	2,155	0	5,476
2	0.087	0	0.210	6,438	561	0	1,352
3	0.105	0	0.225	5,680	598	0	1,276
4	0.141	0.028	0.253	5,383	757	149	1,365
5	0.199	0.095	0.302	5,069	1,007	481	1,533
6	0.285	0.191	0.380	4,541	1,296	868	1,724
7	0.406	0.315	0.498	4,289	1,743	1,350	2,137
8	0.568	0.459	0.676	4,108	2,331	1,886	2,776
9	0.774	0.624	0.924	3,915	3,031	2,444	3,619
10	1.033	0.817	1.248	3,722	3,843	3,042	4,644
Total social groups					**17,324**	**10,220**	**25,902**

593

594

595 **Table 4**

596 Badger social group predictions of the constrained model (pseudo-absences taken from

597 within the potential maximum extent surveyed (see text)) for the Republic of Ireland.

Probability class (deciles)	Estimate density	Estimate lower density	Estimate upper density	Km^2	Total estimate	Lower 95% CI	Upper 95% CI
1	0.087	0	0.253	22,748	1,982	0	5,765
2	0.180	0.079	0.280	6,479	1,165	513	1,817
3	0.203	0.098	0.307	5,868	1,190	576	1,804
4	0.189	0.088	0.291	5,791	1,097	509	1,685
5	0.173	0.083	0.263	5,568	964	463	1,464
6	0.187	0.097	0.277	5,191	971	504	1,438
7	0.265	0.163	0.366	4,884	1,294	797	1,790
8	0.440	0.335	0.544	4,652	2,046	1,559	2,533
9	0.745	0.645	0.846	4,349	3,241	2,804	3,679
10	1.215	1.048	1.381	4,288	5,208	4,495	5,922
Total social groups					**19,159**	**12,221**	**27,898**

598

599

600 **Fig. 1.** The distribution of badger main-sett presence (n = 7,724) locations as used in the
601 national-scale dataset for the Republic of Ireland. The shaded areas represent the
602 potential maximum extent surface (see text for details) from which pseudo-absences were
603 selected during the constrained model development. Inset: The local-scale dataset with
604 the distribution of 278 main setts within the 755km^2 survey area.

605 **Fig. 2.** Relationship between probability values predicted from badger distribution
606 models for the Republic of Ireland (A: non-constrained; B: constrained) and badger main
607 sett density (km^{-2}) within the local-scale area. The points represent the observed density
608 per probability quantile (classes). The solid line is the predicted values; dashed lines are
609 the 95% CI for a prediction (includes both the uncertainty of the mean prediction and the
610 residual).

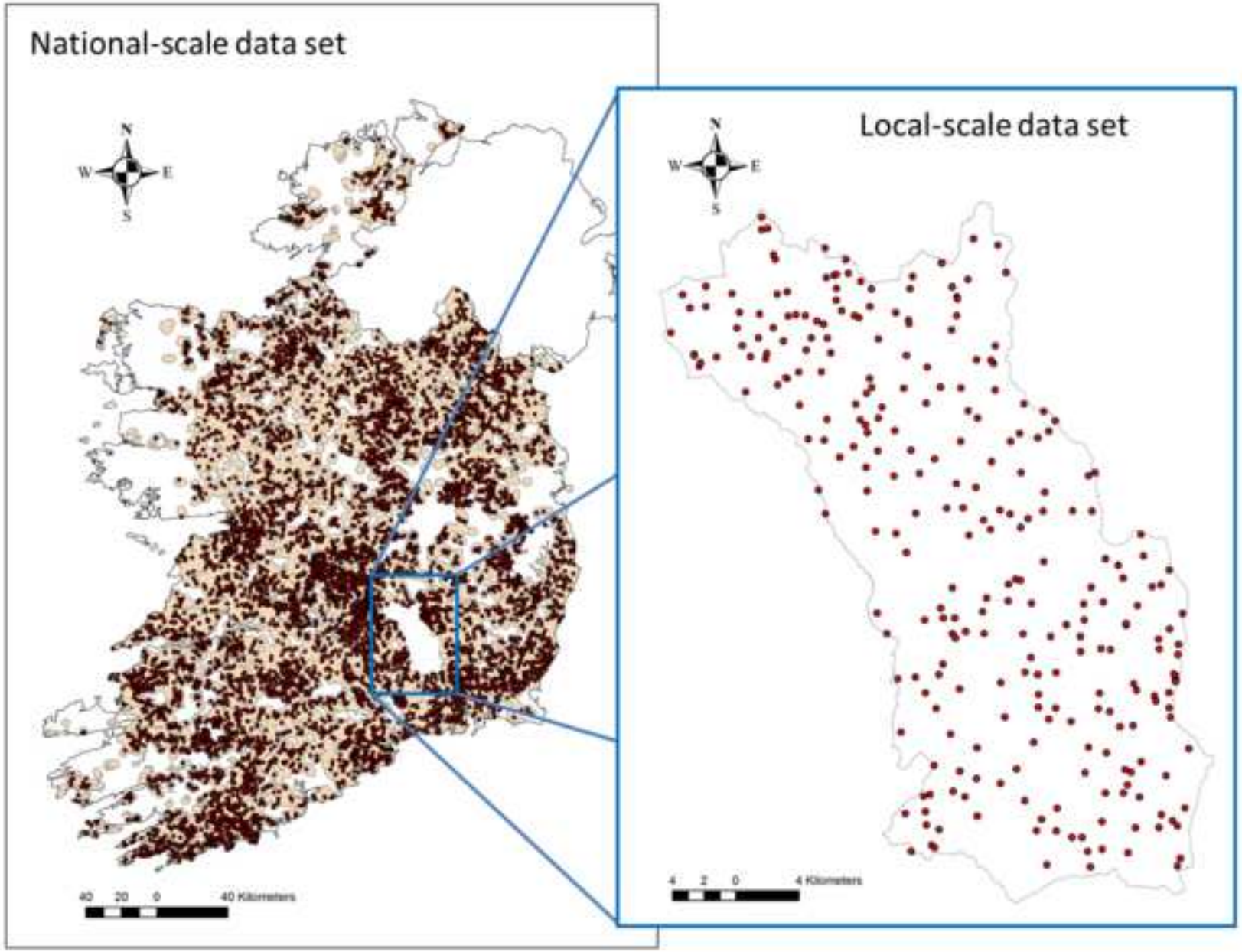
611 **Fig. 3.** Density map of badger social groups (based on main setts km^{-2}) in the Republic of
612 Ireland produced from A: a non-constrained sampling national model; B: a constrained
613 sampling national model.

614 **Fig. 4.** Mean national estimated density of badger social groups in the Republic of
615 Ireland (ROI; present study), Northern Ireland (Feore, 1994; Reid et al., 2012) and
616 England and Wales (Etherington et al., 2009; Judge et al., 2014). A: 95% CI around the
617 national mean. B: The error bars indicate the variation in density across all
618 environmental classes (landscape types).

619

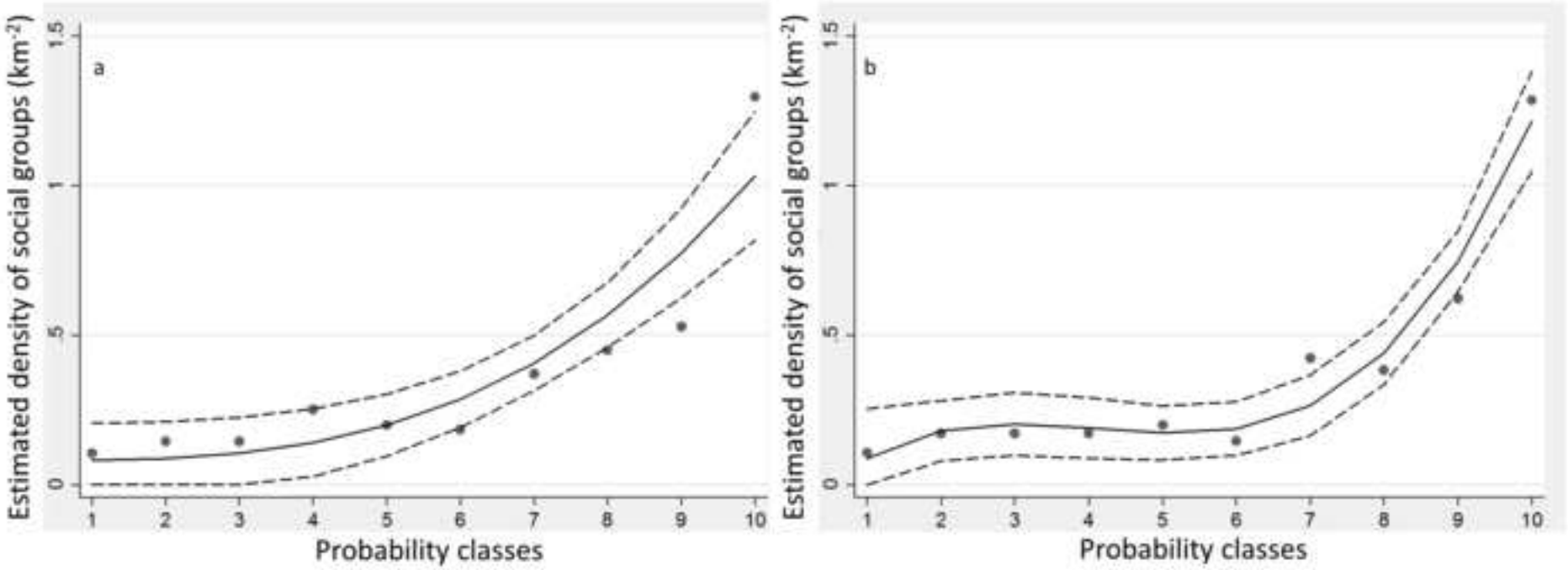
Figure

Click here to download high resolution image



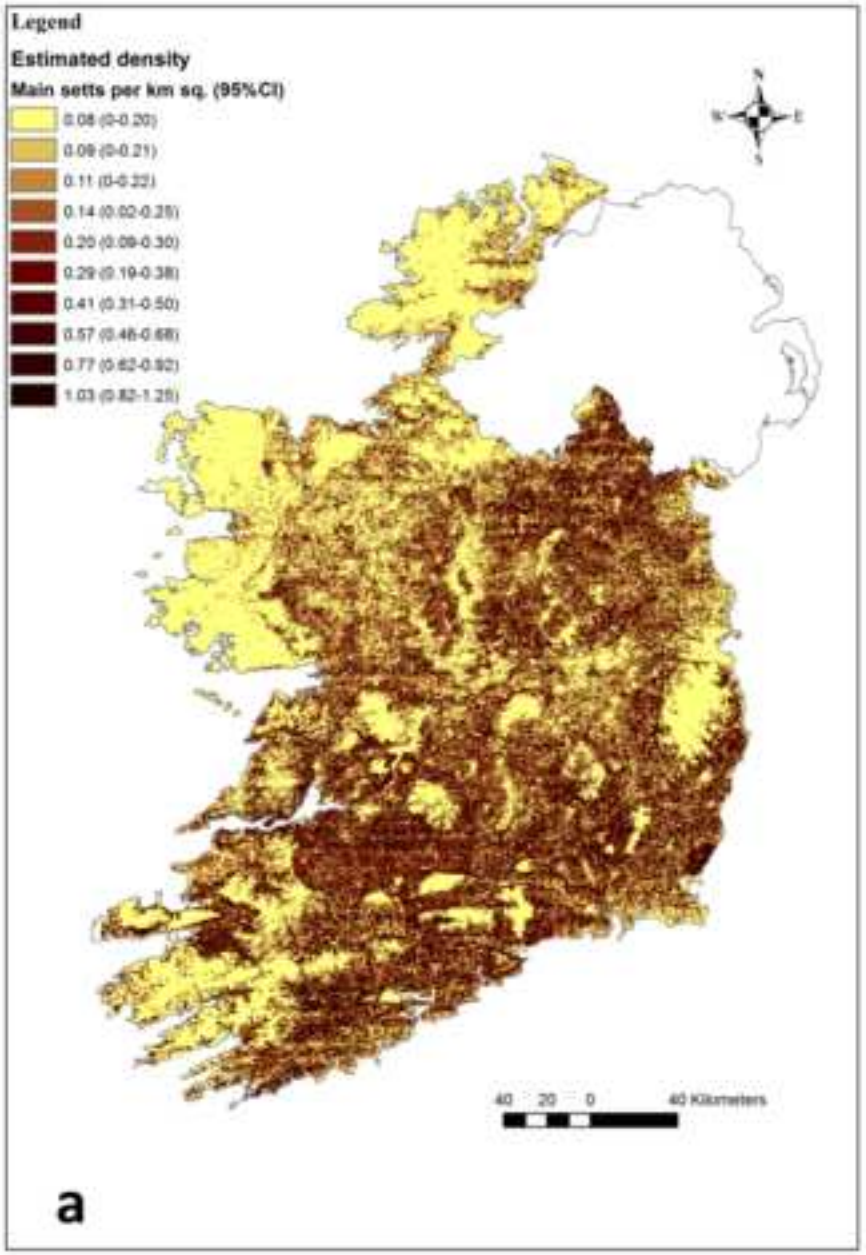
Figure

Click here to download high resolution image



Figure

Click here to download high resolution image



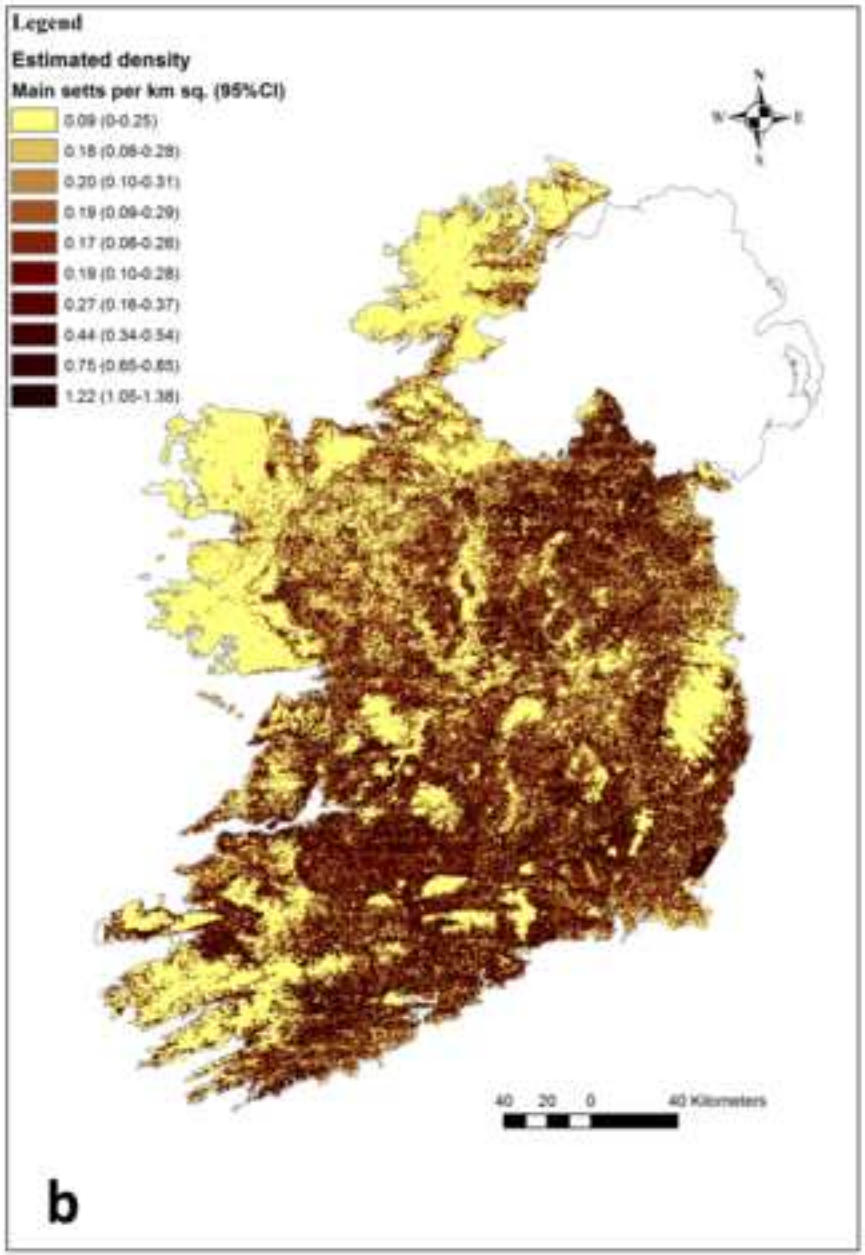
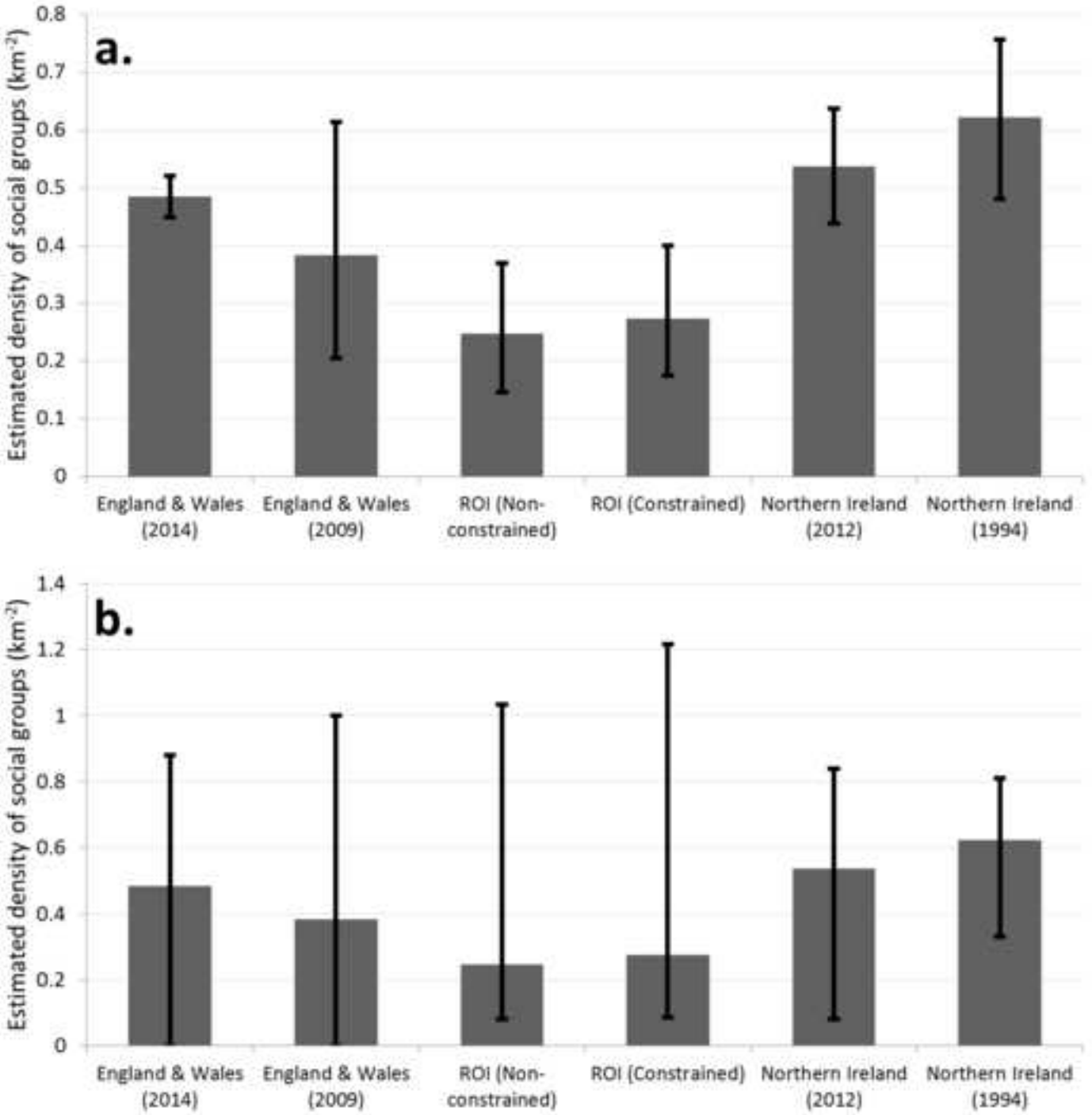


Figure
Click here to download high resolution image



Supplementary material

Appendix S1

Table S1: Final constrained model, with pseudo-absence points drawn from within the potential maximum extent surface (PMES; ~49,000km^2).

Variable		β	SE	z	P>|z|
Elevation (Ele)		-0.002	<0.001	-3.77	<0.001
	Ele^2	<0.001	<0.001	-5.67	<0.001
Slope		0.118	0.013	8.94	<0.001
	Slope^2	-0.004	0.001	-4.86	<0.001
Pasture (300m)		0.001	<0.001	15.73	<0.001
Hedgerow		<0.001	<0.001	26.80	<0.001
Cover		0.079	0.004	18.59	<0.001
Y coordinate		<0.001	<0.001	-4.80	<0.001
X coordinate		<0.001	<0.001	0.42	0.677
Sine (Aspect)		-0.081	0.024	-3.42	0.001
Distance river (DR)		<0.001	<0.001	-7.26	<0.001
	DR^2	<0.001	<0.001	6.44	<0.001
Bog (y/n)		-1.374	0.143	-9.59	<0.001
Made (ground)^ (y/n)		-1.498	0.456	-3.28	0.001
Water edge (y/n)		-1.333	0.321	-4.16	<0.001
Soil type		Wald test (chi^2 (DF: 3) = 36.40			<0.001
Parent material		Wald test (chi^2 (DF: 12) = 173.65			<0.001
Constant		-1.325	0.075	-17.55	<0.001

Table S2: Final non-constrained model, with pseudo-absence points drawn from across the Republic of Ireland (~70,000km^2).

Variable		β	SE	z	P>|z|
Elevation (Ele)		-0.001	<0.001	-1.14	0.255
	Ele^2	<0.001	<0.001	-8.16	<0.001
Slope		0.108	0.014	8.01	<0.001
	Slope^2	-0.004	0.001	-4.63	<0.001
Pasture (300m)		0.001	<0.001	10.31	<0.001
Dry grasslands (300m)		<0.001	<0.001	4.03	<0.001
Hedgerow		0.001	<0.001	31.35	<0.001
Cover		0.312	0.018	16.94	<0.001
	Cover^2	-0.016	0.001	-12.49	<0.001
Y coordinate		<0.001	<0.001	-0.34	0.731
X coordinate		<0.001	<0.001	-6.76	<0.001
Bog (y/n)		-1.484	0.142	-10.47	<0.001
Made (ground)^ (y/n)		-1.602	0.342	-4.68	<0.001
Water edge (y/n)		-1.127	0.325	-3.46	0.001
Soil type		Wald test (chi^2 (DF: 7)) =	143.70		<0.001
Parent material		Wald test (chi^2 (DF:6)) =	55.25		<0.001
Constant		-1.531	0.096	-15.87	<0.001

Table S3: Internal validation of the constrained national model. Models were trained using 70% of the dataset and then predicted to a 30% independent sample. The internal validation procedure was repeated ten times (set 1-10).

	Training (70%)							**Testing (30%)**						
Set	AUC	Cut-point	TSS	Sens	Spec	Kappa	HL-gof	AUC	Cut-point	TSS	Sens	Spec	Kappa	HL-gof
1	0.72	0.4	0.31	64.77	66.08	0.26	0.345	0.72	0.4	0.29	63.75	65.38	0.25	0.244
2	0.72	0.4	0.31	65.45	65.38	0.26	0.155	0.69	0.4	0.27	62.59	64.12	0.24	0.525
3	0.71	0.4	0.30	65.09	65.13	0.26	0.232	0.73	0.4	0.32	65.36	66.95	0.28	0.230
4	0.71	0.4	0.30	64.23	66.2	0.25	0.340	0.73	0.4	0.34	66.34	67.59	0.27	0.378
5	0.71	0.4	0.31	65.34	65.32	0.26	0.186	0.71	0.4	0.30	63.26	66.51	0.25	0.585
6	0.71	0.4	0.30	65.22	65.26	0.26	0.548	0.71	0.4	0.31	64.87	65.95	0.25	0.645
7	0.71	0.4	0.31	63.92	66.73	0.26	0.280	0.71	0.4	0.30	66.77	63.48	0.24	0.331
8	0.71	0.4	0.31	64.38	66.29	0.26	0.153	0.72	0.4	0.30	66.02	64.02	0.27	0.438
9	0.71	0.4	0.31	65.45	65.31	0.26	0.252	0.71	0.4	0.29	64.89	64.48	0.27	0.262
10	0.71	0.4	0.30	64.71	65.74	0.26	0.383	0.71	0.4	0.31	63.43	67.80	0.23	0.167
Mean	0.71	0.4	0.31	64.86	65.74	0.26	0.287	0.71	0.4	0.30	64.73	65.63	0.26	0.380
Max	0.72	0.4	0.31	65.45	66.73	0.26	0.548	0.73	0.4	0.34	66.77	67.8	0.28	0.645
Min	0.71	0.4	0.30	63.92	65.13	0.25	0.153	0.69	0.4	0.27	62.59	63.48	0.23	0.167

AUC: Area Under the ROC Curve; Cut-point: optimum threshold that maximises sensitivity and specificity; TSS: True Skill Statistic; Sens: Sensitivity; Spec: Specificity; Kappa: Cohen's Kappa; HL-gof: Hosmer-Lemeshow goodness of fit test.

Table S4: Internal validation of the non-constrained national model. Models were trained using 70% of the dataset and then predicted to a 30% independent sample. The internal validation procedure was repeated ten times (set 1-10).

	Training (70%)							**Testing (30%)**						
Set	AUC	Cut-point	TSS	Sens	Spec	Kappa	HL-gof	AUC	Cut-point	TSS	Sens	Spec	Kappa	HL-gof
1	0.76	0.5	0.38	67.4	70.89	0.32	0.000	0.78	0.4	0.42	71.95	70.48	0.37	0.200
2	0.77	0.4	0.40	71.23	68.43	0.34	0.000	0.76	0.4	0.39	67.76	70.83	0.33	0.086
3	0.77	0.4	0.40	71.23	69.03	0.34	0.000	0.75	0.4	0.36	69.85	66.40	0.31	0.008
4	0.77	0.4	0.39	70.13	69.35	0.34	0.000	0.76	0.5	0.37	66.47	70.05	0.33	0.001
5	0.76	0.4	0.39	70.35	68.6	0.33	0.000	0.77	0.5	0.40	68.28	71.43	0.34	0.018
6	0.76	0.4	0.39	70.61	68.81	0.33	0.000	0.77	0.4	0.39	71.35	68.05	0.33	0.057
7	0.77	0.4	0.40	71.19	68.78	0.34	0.000	0.75	0.4	0.36	69.36	66.56	0.33	0.000
8	0.77	0.4	0.39	71.22	68.22	0.34	0.000	0.76	0.5	0.40	68.99	71.04	0.34	0.002
9	0.76	0.4	0.39	70.82	68.24	0.33	0.000	0.77	0.5	0.40	68.26	71.95	0.35	0.039
10	0.76	0.4	0.39	70.7	68.38	0.33	0.000	0.78	0.4	0.41	71.46	69.30	0.36	0.007
Mean	0.77	0.4	0.39	70.49	68.87	0.33	0.000	0.77	0.4	0.39	69.37	69.61	0.34	0.042
Max	0.77	0.5	0.40	71.23	70.89	0.34	0.000	0.78	0.5	0.42	71.95	71.95	0.37	0.200
Min	0.76	0.4	0.38	67.4	68.22	0.32	0.000	0.75	0.4	0.36	66.47	66.4	0.31	0.000

AUC: Area Under the ROC Curve; Cut-point: optimum threshold that maximises sensitivity and specificity; TSS: True Skill Statistic; Sens: Sensitivity; Spec: Specificity; Kappa: Cohen's Kappa; HL-gof: Hosmer-Lemeshow goodness of fit test.

Appendix S2

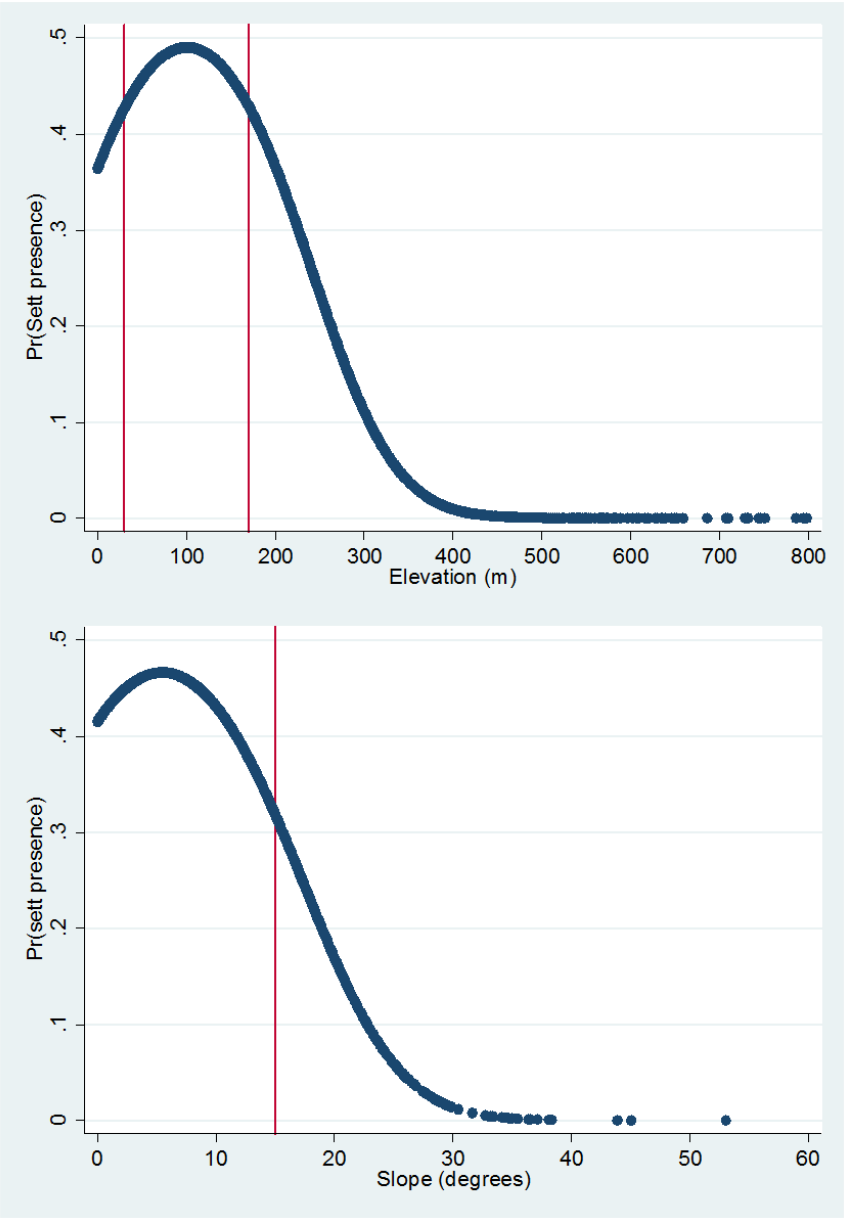


Figure S1. An example of the quadratic relationship between the predicted probability of sett occurrence and elevation (above) and slope (below). Setts are most likely constructed at 50-170m above sea level and in moderate slopes below 15º (red reference lines).